

Tõenäosuse ja statistika algkursus

Praktikum 4

Ülesanne 1

Katsetulemuse X jaotust on kirjeldatud alltoodud tõenäosusfunktsiooni abil.

x	-1	0,1	0,3	2
$P(X=x)$	0,1	0,3	0,1	0,5

- a) Kui tõenäoliselt omandab ülaltoodud jaotusega juhuslik suurus väärtuse 0,2?
- b) Leia uuritava tunnuse keskväärus EX .

Ülesanne 2

Viskame kuut täringut. Millise jaotusega on viske tulemusel saadud viite arv (millisesse jaotuste perre antud jaotus kuulub ja millise konkreetse jaotusega on tegemist)? Kuidas seda informatsiooni jaotuse kohta saab kirja panna?

Kui tõenäoliselt omandab kirjeldatud juhuslik suurus väärtuse 1? Kui tõenäoliselt väärtuse 2?

Ülesanne 3

Viskame ausat münti 10 korda ja loeme kokku kümne viske peale saadud kullide arvu (tähistame saadud kullide arvu X -ga). Millise jaotusega on X ? Leia tõenäosus $P(X=5)$.

Ülesanne 4

- a) Linnas, kus elab 10000 inimest, on üks kiirabiauto, mis vajadusel annab esmaabi ja viib abivajaja Tartusse (Tartu on aga kaugel ja sõidu peale kulub terve päev). Tõenäosus, et üks linnakodanik vajab haiglaabi ühel konkreetselt päeval on $1/20000$. Kui suur on tõenäosus, et ühest kiirabiautost linnale ei piisa (haiglaravi vajavate inimeste arv päevas on suurem kui 1)?
- b) Oletame, et linnas elab hoopis 20000 inimest, aga üks elanik vajab haiglaravi tõenäosusega $1/40000$. Leia tõenäosused, et päevas vajab haiglaravi 0 või 1 inimest. Võrdle neid eelmises punktis leitud tulemustega.
- c) Me ei tea linnakese suurust, aga teame, et keskmiselt sattub iga päev selles linnas haiglaravile 0,5 inimest. Leia tõenäosused, et päevas vajab haiglaravi 0 või 1 inimest. Võrdle neid eelmistes punktides leitud tulemustega.

Ülesanne 5

Eestis leiab keskmiselt aset 7 infarkti päevas. Ühe päeva jooksul toimuvate infarktide arvu jaotus on ligilähedaselt Poissoni jaotusega juhuslik suurus. Kui tõenäoliselt juhtub ühel päeval: (a) 0 infarkti (mitte ühtegi); (b) 1 infarkt; (c) 7 infarkti?

Ülesanne 6

Kui soovitakse teada saada inimese genoomi, siis loetakse (sekveneeritakse) juhuslikult valitud alguskohast mingi arv tähti ehk aluspaare (näiteks loetakse järjest 250 tähte). Seejärel valitakse uus juhuslik koht genoomist ja loetakse taas 250 tähte pikkune lõik (lugem). Neid lugemeid tehakse väga palju ja hiljem üritatakse neist lugemitest nagu piltmõistatuse tükkidest terve inimese genoom kokku panna.

Kui iga koht inimese genoomis on keskmiselt kaetud kümne lugemiga, siis räägitakse, et inimest on sekveneeritud keskmise katvusega 10 (ühte kohta genoomis on loetud 10 korda). Lugemite arv, mis ühte positsiooni katab, on ligilähedaselt Poissoni jaotusega juhuslik suurus – vaata ka kõrval olevat väljavõtet Illumina sekvenaatoriga kaasaskäivast dokumentatsioonist

2. Estimating the number of times a base is expected to be sequenced.
Lander and Waterman made two assumptions about the sequencing:

- Reads will be distributed randomly across the genome
- Overlap detection doesn't vary between reads.

Based upon these two assumptions, they reached the conclusion that the number of times a base is sequenced follows a Poisson distribution. The Poisson distribution can be used to model any discrete occurrence given an

(https://www.illumina.com/documents/products/technotes/technote_coverage_calculation.pdf).

Kui tõenäoliselt me ühte konkreetset kohta genoomis ei näe üheski lugemis (ei loe kordagi)?

Mõtlemiseks: Inimgenoomi pikkus on umbes 3 000 000 000 tähte. Mitu tähte jääb meil ootuspäraselt nägemata (ei loeta kordagi) kui keskmiselt loeme (sekveneerime) iga tähte (iga positsiooni genoomis) 10 korda?

Lisaülesanne – pane ennast proovile!

(kõik järgnevad ülesanded on lahendatavad binoomjaotust kasutades – vaata, kas oskad binoomjaotust rakendada päriselust pärit ülesannetes!)

Teame ühe ohtliku bakteri genoomist 60 tähte pikka DNA lõiku (DNA koosneb tähtedest A,C,T,G), mis esineb teadaolevalt ainult ohtlikus bakteris ja mitte üheski teises elusolendis.

Loeme (sekveneerime) proovis olevat bakterit (oletame, et saaksime kogu bakteri genoomi korraga lugeda) ja vaatame, kas leiame loetud tähejadas meie poolt otsitavat 60 tähe pikkuse lõigu. Kui leiame, siis loeme ohtliku bakterit leituks (test osutub positiivseks).

Paraku teeb tänapäevane tehnoloogia DNA lugemisel tõenäosusega 0,01 tähevea (lugemisvead esinevad teineteisest sõltumatult).

- a) Mitu korda peaksime proovis oleva bakteri genoomi lugema, kui sooviksime, et avastaksime ohtliku bakterit (kui ta esineb proovis) vähemalt tõenäosusega 0,99?

Oletame, et proovis võib esineda ka sarnane ohutu bakter, millel on peaaegu samasugune 60 tähte pikk DNA lõik – vaid 1 täht kogu lõigu kohta on erinev.

- b) Milline on tõenäosus anda ekslikult häiret, kui proovis on ohtliku bakteri asemel tema ohutu sugulane (ja me üritame tagada avastamistõenäosust 0,99 - loeme proovis oleva bakteri genoomi niimitu korda kui on minimaalselt vaja avastamistõenäosuse 0,99 saavutamiseks)?
- c) Kuidas saavutada ohtliku bakteri avastamistõenäosus 0,99 (kui ohtlik bakter tegelikult on proovis) ja samal ajal tagada, et ainult ohutu sugulasliigi esinemise korral proovis oleks valepositiivse testitulemuse tõenäosus väiksem kui 0,01?

Lisaküsimuse vastused (ära kohe piilu – proovi enne ülesannet ise lahendada!):

- a) 6 või enam korda
b) 0,019834...
c) Lugesdes tundmatu bakteri genoomi 9 korda ja nõudes, et kahel või enamal korral näeksime meie poolt otsitud 60-tähe pikkust DNA-lõiku. Kui proovis olev bakter oleks ohtlik bakter, siis tuvastaksime ta umbes ~99,05% tõenäosusega, kui aga proovis oleks ohutu sugulasliik siis tõenäosusega ~0,0004 peaksime teda ohtlikuks bakteriks. Lahenduseks sobivaid kombinatsioone on teisiigi (näiteks nõuda, et proovi 12 kordse lugemisel näeksime 3 või enamal korral ohtliku bakteri signatuuri).